Federal Aviation Administration

DOT/FAA/AM-10/2
Office of Aerospace Medicine
Washington, DC 20591

Examining Smoking-Induced Differential Gene Expression Changes in Buccal Mucosa

Doris M. Kupfer[1]
Vicky L. White[1]
Marita C. Jenkins[2]
Dennis Burian[1]

[1]Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

[2]Advancia, Inc.
Oklahoma City, OK 73104

January 2010

Final Report

OK-10-0077-JAH

NOTICE

____

This publication and all Office of Aerospace Medicine technical reports are available in full-text from the Civil Aerospace Medical Institute's publications Web site:
www.faa.gov/library/reports/medical/oamtechreports

Technical Report Documentation Page

1. Report No.
DOT/FAA/AM-10/2

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Examining Smoking-Induced Differential Gene Expression Changes in Buccal Mucosa

5. Report Date
January 2010

6. Performing Organization Code

7. Author(s)
Kupfer DM,[1] White VL,[1] Jenkins MC,[2] Burian D[1]

8. Performing Organization Report No.

9. Performing Organization Name and Address

[1]FAA Civil Aerospace Medical Institute
P.O. Box 25082
Oklahoma City, OK 73125

[2]Advancia, Inc.
Oklahoma City, OK 73104

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

12. Sponsoring Agency name and Address

Office of Aerospace Medicine
Federal Aviation Administration
800 Independence Ave., S.W.
Washington, DC 20591

13. Type of Report and Period Covered

14. Sponsoring Agency Code

15. Supplemental Notes

Work was accomplished under approved task AM-TOXLAB

16. Abstract

Gene expression changes resulting from conditions such as disease, environmental stimuli, and drug use can be monitored in the blood. However, a less invasive method of sample collection is of interest because of the discomfort and specialized personnel necessary for blood sampling, especially if multiple samples are being collected. Buccal mucosa (cheek swabs) are easily collected and may be an alternative sample material for biomarker testing. A limited number of studies, primarily in the smoker/oral cancer literature, address this tissue's efficacy as an RNA source for expression analysis. The current study was undertaken to determine if total RNA isolated from buccal mucosa could be used as an alternative tissue source to assay relative gene expression. In this study, qPCR and microarray analyses were used to evaluate gene expression in buccal cells. Initially, qPCR was used to assess relative transcript levels of four genes from whole blood and buccal cells collected from the same seven individuals at the same time. The RNA isolated from buccal cells was degraded but was of sufficient quality to be used with RT-qPCR to detect expression of specific genes. Second, buccal cell RNA was used for microarray-based differential gene expression studies by comparing gene expression between smokers and nonsmokers. An amplification protocol allowed use of 150-fold less buccal cell RNA than had been reported previously with human microarrays. We report here the finding of a small number of statistically significant differentially expressed genes between smokers and nonsmokers, using buccal cells as starting material. Gene Set Enrichment Analysis confirmed that these genes had a similar expression pattern as results from another study. Our results suggest that despite a high degree of degradation, RNA from buccal cells from cheek mucosa could be used to detect differential gene expression between smokers and nonsmokers. However, the RNA degradation, increase in sample variability, and microarray failure rate show that buccal samples should be used with caution as source material in expression studies.

17. Key Words
Buccal Cells, Smoking, Differential Gene Expression, qPCR, Microarray

18. Distribution Statement
Document is available to the public through the Defense Technical Information Center, Ft. Belvoir, VA 22060; and the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	27	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

AUTHORS' CONTRIBUTIONS

DMK participated in design of study, data analysis, and drafted the manuscript; VLW isolated RNA and performed amplifications; MCJ performed qPCR and microarray hybridizations; DB conceived the study and reviewed the manuscript.

ACKNOWLEDGMENTS

Our thanks to Kristina Holton and Dana Farber, Cancer Research Institute, Boston, MA, for assistance on analysis approaches, especially Rank Product.

CONTENTS

Examining Smoking-Induced Differential Gene Expression Changes in Buccal Mucosa

BACKGROUND

Blood has been shown to be a responsive tissue that is useful for monitoring gene expression changes due to disease, environmental, biological or drug effects. However, for studies performed in human subjects, a less invasive tissue source for biomarker monitoring is of interest due to the discomfort, required skill level, and cost of blood collection, especially for repeated-measures studies. Buccal mucosa (from cheek swabs) is an easily accessed tissue and has been used successfully to obtain DNA for genotyping studies (1). However, the literature is limited as to the usefulness of RNA from buccal cells as a substrate for gene expression testing, presumably due to concern regarding high concentrations of RNases in saliva which are known to rapidly degrade RNA in these cells (2). qPCR has been used to detect expression changes in genes from the P450 family using snap frozen surgical samples (3) and from brushed exfoliated buccal cells (4, 5). These studies suggested that buccal cells might serve as an alternative to blood in qPCR assays examining gene expression profiles after exposure to environmental toxins, tobacco smoke, drugs, nutrients, or the presence of certain cancers. With RNA purified from brushed exfoliated buccal cells, Sridhar et al. (6) used microarrays to compare expression levels between smokers and nonsmokers, and to compare expression patterns between buccal cells and bronchial epithelium in smokers and nonsmokers (7) by Gene Set Enrichment Analysis (GSEA) (8).To our knowledge, buccal cells have not been used with a whole transcriptome approach to investigate differential gene expression. A successful study of this type would suggest that buccal cells have efficacy as source material for biomarker discovery or in a gene expression monitoring system.

We describe here both qPCR and microarray approaches. The qPCR study used matched blood and brushed buccal samples from the same subjects. Relative expression levels of four genes allowed comparison of tissue sources and subject differences. RNA from buccal cells was highly degraded; nonetheless, expression could be detected by qPCR for all four transcripts tested. This was sufficient evidence of the potential of buccal cells to follow up on the work of Sridhar et al. (6) and use microarrays for differential gene expression analysis on the transcriptome level in smokers and nonsmokers. An important consideration was the availability of the Smoking Induced Epithelial Gene Expression Database, (SEIGE) (7) and smoker buccal mucosa-specific gene lists (6), against which results from this study could be compared to confirm our method.

Our data were first analyzed for differences between smokers and nonsmokers using Significance Analysis of Microarray (9) and Rank Product (10) for detection of significant gene expression differences between the smoker and the nonsmokers in our study. These analyses resulted in a list of candidate marker genes from each method. Ingenuity Pathway Analysis (11) was used to find functional networks containing the differentially expressed genes. The gene lists were also examined for transcriptional coregulation by searching the promoters of differentially expressed genes for transcription factor binding sites (TFBS) using PAINT (12) to access the TRANSFAC database of known TFBS. Specifically, we identified 103 genes with Rank Product analysis that had increased expression in smokers. Pathway analysis showed five function networks involving 91 of the 103 target genes. Network functions included cell cycle, cell growth, proliferation and movement, gene expression, and immunological disease. Upstream sequence analysis showed 41 target genes containing binding sites for at least one of three widely expressed transcription factors. Twenty-five genes were identified using SAM analysis. Similar to the RP results, 13 of these genes fell into one of two functional networks that had shared roles in tumor morphology, metabolic disease, lipid and carbohydrate metabolism, and which contained binding sites for at least one of two widely expressed transcription factors. These results suggest that many of these genes are co-regulated and that the transcriptional response affects numerous cellular functions.

Both gene lists were further analyzed using GSEA to compare the buccal dataset against the Sridhar gene sets. The comparisons showed that the genes in the published sets changed expression in the same direction in our buccal array data.

The results of the study suggest that buccal mucosa may indeed be useful for factors selected carefully for optimum expression change in buccal tissue. However, the extensive random degradation, which may vary between subjects, suggests a loss of sensitivity and possibly the need for multiple sampling, which is costly. It also suggests that due to the extensive degradation found, it seems unlikely to be a reliable source for biomarker discovery.

RESULTS

Quality Assessment

Initially, we determined the quality of RNA purified from buccal mucosa. Matched blood and buccal total RNAs from seven subjects were purified (see Materials and Methods). RNA quality was assessed on the Agilent Bioanalyzer RNA using Nano 6000 chips (Figure1). Buccal RNA samples were found to be severely degraded with RNA Integrity Numbers (RINs) routinely less than three. In contrast, RINs from the blood samples were greater than seven in all cases (Table 1). These results indicate that the buccal RNA was not of high quality.

qPCR Validation

To determine if RNA from buccal cells could be useful for marker analysis, we chose to perform qPCR on these paired samples. To test whether RNA degradation was non-specific or directional, we chose an amplification method that primed reverse transcription from random primed hexamers. Primers to four genes were used: ITGA5, ANKRD28, TMEM8, and RPS3A. BioGPS (13) values for these four genes indicated an approximate expected ratio of buccal cells (salivary gland used for estimate) versus blood (Table 2). Primers were made to the 3 prime (3') ends of all four genes. To determine whether RNA degradation was random or specific by gene region, primers to upstream regions of ITGA5, ANKRD28 and TMEM8 were also designed (Table 3).

The WT Ovation Pico kit was used for amplification of all 14 samples, both blood and buccal, and the subsequent product used for qPCR with the primer pairs detailed above.

An average over the seven subjects showed that there was a lower apparent transcript copy-number for each tested gene in buccal mucosa RNA than in blood RNA. In some subjects, no Ct was calculated, and the differences between apparent transcript levels were greater than the mean value indicates. As seen from the increased standard deviations, RNA from buccal cells had greater variability in Cts, suggesting that buccal RNA quality is also more variable than blood RNA (Table 4).

When specificity of degradation was investigated, no clear pattern was evident. ITGA5 showed a 32-fold difference from 5' to 3' in buccal mucosa compared to an approximately three-fold difference in blood, but most reactions with ITGA5 primers with buccal RNA failed. ANKRD28 showed no change in 5'/3' ratio in either RNA source. TMEM8 showed an increase in 3' signal over 5' signal in buccal mucosa but a 3' preferential loss of target in blood. Due to the short transcript length of RPS3A, no 5' primer set was designed. This initial analysis of the quality of buccal RNA shows that, in general, there were lower but detectable levels of target mRNA in buccal mucosa when compared to blood (Table 4). These results do not differentiate between tissue-specific expression differences or degradation; however, when the expression data from BioGPS and the RINs were factored into our analysis, the differences in Cts were greater than expected from expression data and likely due to degradation. The variability of results from buccal cells suggests that the degradation seen in the buccal samples is not predictably occurring in a directional fashion but randomly such that transcript size has no effect.

The reduced signal detected in buccal versus blood samples with the WT amplification method led us to hypothesize that a 3'-specific amplification of sample assayed with 3'-specific primers would increase signal-to-noise ratios and have increased sensitivity in expression assays. To investigate this possibility, the same samples were amplified with the Ovation RNA Amplification System V2, a 3' specific method. Table 5 shows a comparison of the amplification results using the 3' targeted primers and both buccal mucosa and blood derived RNA template. For all three genes, 3' amplification resulted in a Ct decrease, i.e., an apparent increase in copy number, although Cts from buccal mucosa RNA tested with primers to ITGA5, remained greater than 31. Relative Cts from ANKRD28 and TMEM8 between buccal RNA and blood RNA compare favourably with data from BioGPS comparing salivary gland to whole blood. However, ITGA5 values did not correspond particularly well, suggesting that ITGA5 was more sensitive to degradation than the other genes tested.

Microarray Study

Our ability to detect expression of genes by qPCR, most at levels well above background, in 3' amplified samples lead us to hypothesize that buccal samples could be used for differential expression testing by microarray analysis. Amplification of buccal sample RNA has the advantage of not requiring repeated sample collection and/or pooling of material from multiple collections. The previous work of others (5, 6) led to the further hypothesis that a comparison of smokers and nonsmokers was a model system likely to allow detection of differentially expressed genes. Affymetrix Human U133 plus 2.0 arrays were used for a global evaluation of gene expression changes between four smokers and four nonsmokers. All female subjects were used to prevent any gender bias in the data, and both cheeks from each subject were sampled. Total RNA was isolated and evaluated for quality as for the qPCR samples. One cheek sample from each subject was arbitrarily assigned to one of two groups, a or b (Materials

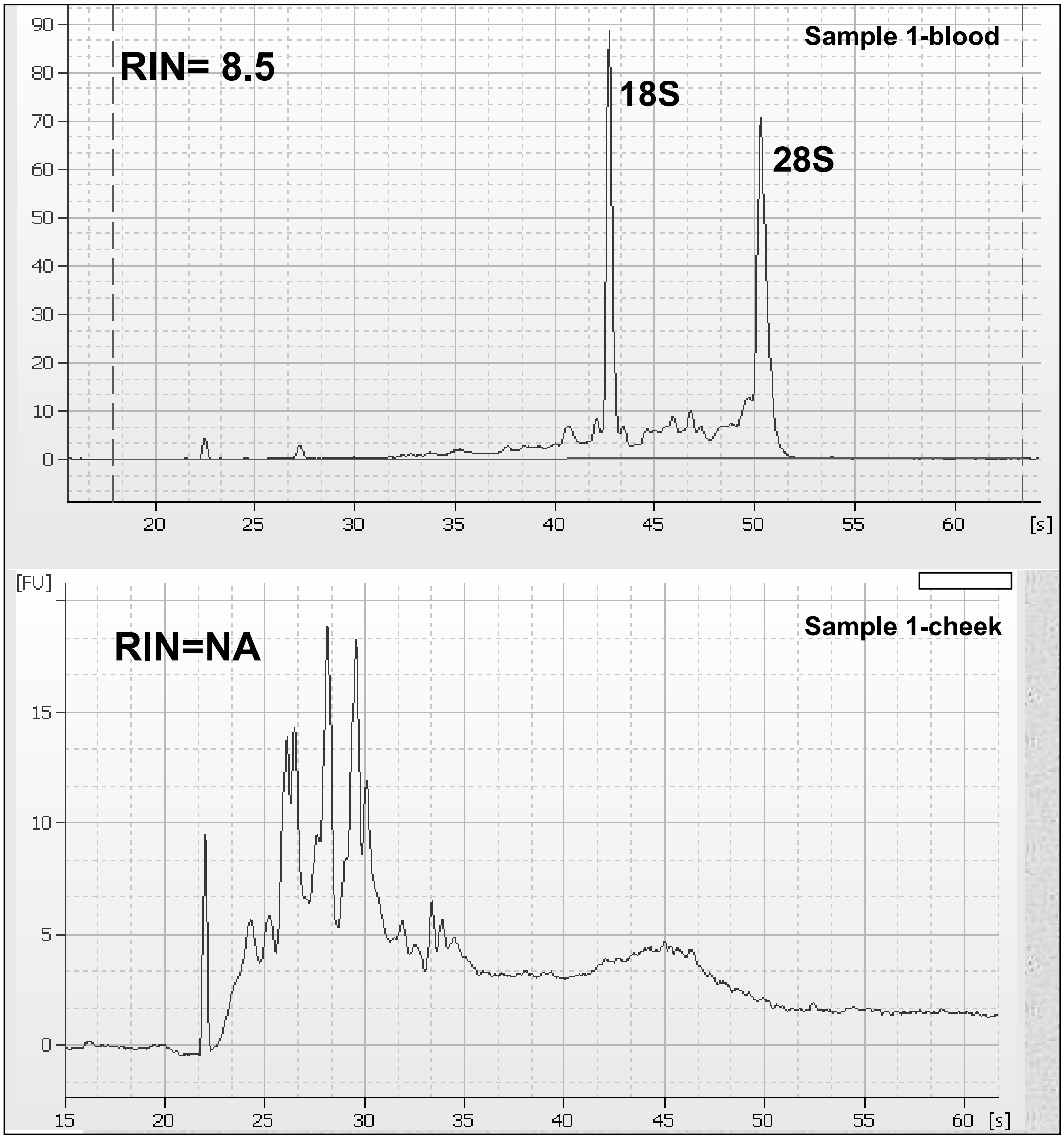


Figure 1. Representative qPCR matched blood and buccal mucosa samples. The buccal RNA (cheek) appears to be heavily degraded compared to the blood RNA since there is no evidence of 18 or 28S rRNA peaks and the bulk of material is migrating rapidly, indicating small size. RIN, RNA integrity number. NA, No RIN could be determined.

Table 1. Total RNA yield from blood and buccal samples used in study.

qPCR Samples	RNA Conc. ng/uL	Vol ul	Total ng	RIN
1-Cheek	1.45	30	43.5	2.5
2-Cheek	3.12	30	93.6	N/A
3-Cheek	3.22	30	96.6	N/A
4-Cheek	4.57	30	137.1	2.5
5-Cheek	5.07	30	152.1	2.6
6-Cheek	5.63	30	168.9	2.7
7-Cheek	3.99	30	119.7	N/A
1-Blood	149.76	80	11980.8	8.5
2-Blood	71.83	80	5746.4	8.7
3-Blood	51.38	80	4110.4	8.1
4-Blood	34.15	80	2732.0	7.4
5-Blood	90.19	80	7215.2	8.0
6-Blood	388.5	80	31080.0	7.8
7-Blood	130.93	80	10474.4	8.1

Microarray Samples	RNA ng/uL	Vol ul	Total ng	RIN
NS21a	13.94	30	418.2	2.8
NS22a	6.43	30	192.9	2.5
NS23a	9.12	30	273.6	N/A
NS24a	13.23	30	396.9	N/A
Sm25a	2.90	18	52.2	N/A
Sm26a	6.55	30	196.5	N/A
Sm27a	9.20	30	276.0	2.3
Sm28a	9.30	30	279.0	2.5
11Sma	2.93	30	87.9	ND
12NSa	10.15	30	304.5	ND
NS21b	3.91	30	117.3	N/A
NS22b	25.43	30	762.9	2.3
NS23b	7.19	30	215.7	N/A
NS24b	12.90	30	387.0	N/A
Sm25b	6.18	30	185.4	N/A
Sm26b	13.46	30	403.8	2.1
Sm27b	24.43	30	732.9	N/A
Sm28b	11.79	30	353.7	N/A
11Smb	5.99	30	179.7	ND
12NSb	11.51	30	345.3	ND

N/A No RIN could be calculated
ND Not done

Table 2. BioGPS approximate expression values for blood and salivary gland for the genes tested via qPCR.

	Ankrd28	Tmem8	Rps3a	Itga5
Salivary gland	130	2000	60,000	2000
Blood	130	7000	>100,000	7000

Values are approximate signal strength values from BioGPS Human U133A gcRMA dataset, accessed 2009-07-16 [13].

Table 3. Primers used for qPCR portion of study.

Gene	Refseq Accession ID	Affymetrix Probeset ID	Transcript Length (bp)	Sense Primer	Anti-sense Primer	Position 5' (bp)	Cycling Temperature degC	Concentration (nM) sense, anti-sense	% Efficiency	Amplicon Size(bp)
ANKRD28	NM_015199	226025_at	6339	CAGATGATTTTGGC AGGACTTG	GTTGGCAGCAGCGTAGTG	1345	50	250,150	96	143
				TGCTGAGATGTTAA TTGATAC	TCCACAGAATTGACTTGAG	2552	56	300,300	94	150
				CATAAATTAAGCATCACTAAAGTCTC	CCGAAATATCAGCCTTCTCTC	4379	54	300,250	91	199
ITGA5	NM_002205	201389_at	4267	GCTGGACTGTGGA GAAGAC	AAGTGAGGTTCAGGGCATTC	1994	56	300,250	95	107
				TCAGACATTGGCACCTAATC	TTCCTGGCTTCTCCTAAATC	3879	54	300,300	94	104
TMEM8	NM_021259	222718_at	2561	CTTCCAGAGGTTTCTCATAC	GCGTCATACTTCTCATCC	584	57	250,300	95	118
				CGTCAGCAGAATGTATGTG	TCCTGGTTCCATAGTCTCC	1372	54	300,300	98	110
				CCCTGCCTCTTTGCCTTC	TAGTAGTAGTTGTCGCTAGTCATC	2067	57	250,250	95	179
RPS3A	NM_001006	200099_s_at	930	CACCAGGACCCAAGGAACC	CCAACACAGAACAGACGAAGC	198	60	200,200	94	276

Table 4. qPCR results comparing blood and buccal RNA across four genes with whole transcriptome amplified template.

	Tmem8 2067*	Tmem8 1372	Tmem8 584	Ankrd 4387	Ankrd 2552	Ankrd 1345	Itga5 3879	Itga5 1994	RPS3A 198
Cycle temp	57deg	57deg	54deg	54 deg	56 deg	50 deg	54 deg	56 deg	60 deg
Product	179bp	110bp	118bp	199bp	150bp	143bp	104bp	107bp	276bp
Subject									
Buccal 1	33.76	No Ct	34.93	No Ct	30.33	28.91	31.7	No Ct	31.86
Buccal 2	29.72	29.15	26.83	31.64	29.52	27.47	No Ct	No Ct	28.65
Buccal 3	28.18	30.52	29.27	30.81	35.73	30.13	No Ct	38.64	28.49
Buccal 4	24.96	27.74	28.09	31.8	33.55	31.98	35.71	No Ct	26.82
Buccal 5	25.31	26.56	26.39	31.36	32.8	34.02	No Ct	No Ct	27.38
Buccal 6	35.22	34.85	34.1	No Ct	No Ct	37.42	No Ct	No Ct	ND
Buccal 7	29.05	29.24	35.31	No Ct	No Ct	No Ct	No Ct	No Ct	33.19
Mean	29.46	29.68	30.7	31.4	32.39	31.66	33.71	38.64	29.4
StDev	3.61	2.63	3.65	0.38	2.24	3.32	2.01	0	2.33
Blood 1	27.03	26.49	25.78	30.53	29.28	30.36	24.86	26.67	26.31
Blood 2	28.5	28.28	26.58	29.04	28.87	28.86	24.41	26.42	26.66
Blood 3	28.08	27.71	26.7	29.03	29.93	28.33	24.28	25.34	26.1
Blood 4	27.64	26.53	26.57	29.43	29.61	29.68	25.14	25.62	27.11
Blood 5	27.3	27.63	25.32	29.82	28.71	28.89	24.72	25.29	26
Blood 6	27.96	27.67	26.02	29.68	29.09	29.45	26.69	28.92	ND
Blood 7	27.8	28.48	27.62	29.49	28.73	28.06	24.34	27.14	26.82
Mean	27.76	27.54	26.37	29.57	29.17	29.09	24.92	26.49	26.5
StDev	0.46	0.78	0.69	0.48	0.43	0.74	0.78	1.19	0.4
mRNA size	2529 bp	2529bp	2529bp	6339bp	6339bp	6339bp	4248bp	4248bp	903bp
Controls^									
NTC	39.4	No Ct	No Ct	No Ct	No Ct	No Ct	No Ct	No Ct	No Ct
WT amplification	28.4	27.28	26.45	29.47	29.23	29.48	24.63	26.59	25.94
3' amplification	26			22.9			24.4		20.5

Results are in Ct values

* Gene names are followed by 5' position of amplification on mRNA.

^WT and 3' Controls are RNA from pooled blood samples

NTC, No template control ND, not done Bp, basepair Deg, degree centigrade

StDev, Standard deviation

Table 5. qPCR results comparing methods of template amplification.

Template	Itga5b 3'	Itga5b WT	Tmem8 3'	Tmem8 WT	Ankrd28 3'	Ankrd28 WT
Buccal 4	31.25	39.15	20.38	24.96	20.68	31.8
Buccal 5	35.81	No Ct	20.36	25.31	20.92	31.36
Blood 4	21.43	23.96	21.87	30.68	20.33	29.43
Blood 5	20.56	23.9	21.11	30.23	20.22	29.82
Control	21.18	23.8	21.63	28.4	20.23	29.47

All amplifications were performed using Nugen Kits, see Materials and Methods

3'- RNA 3' amplified via a poly T primer

WT- RNA whole transcriptome amplified with random hexamers and poly T primers

Control is a pooled sample from blood, see Materials and Methods

and Methods). Figure 2 shows the BioAnalyzer traces from all 16 samples with a trace representative of the quality of RNA usually purified from blood. As seen with the samples used in the qPCR study, the samples show no evidence of rRNA peaks and a range of degradation product sizes; an RIN be calculated in only a third of the samples could.

Quality Assessment of the Arrays

Following hybridization, each array was examined for quality. Table 6 lists the percent present (%p) and scaling factor (SF) values determined using the Gene Chip Operating Software (Affymetrix, Inc.; Materials and Methods). Two arrays, NS21a and Sm27a, had remarkably low %p and especially high SFs, both indicators of arrays that are suspect for data quality. Additionally, the same two arrays had much lower signal intensities (Figure3). The normalized unscaled standard error (NUSE) (14) calculations had high median values and large interquartile range for these two arrays (Table 6). Samples from the same subject's opposite cheek did not show the same set of quality control issues, further evidence that RNA quality from buccal cells is inconsistent. Neither sample could have been predicted to be of lesser quality from the BioAnalyzer traces (Figure2A). Due to the poor quality of these two arrays, they were removed from further analysis. Two other arrays, Sm28a and b, had elevated NUSE parameters compared to other subjects but did not have a low %p or high SF, and so were not removed as the observed differences were likely subject-dependent and were more likely due to biological diversity between subjects.

Microarray Data Analysis for Differential Expression

A study using Affymetrix hgU133A arrays to compare gene expression in smokers and never-smokers using RNA from buccal mucosa and nasal swabs was published by Sridhar et al. (6). This group performed an extensive microarray analysis of gene expression in bronchial lavage samples from smokers, former-smokers, and never-smokers and developed a list of 314 genes differentially expressed in smokers in this tissue (7, 15). Using Gene Set Enrichment Analysis (GSEA), Sridhar and associates examined the smoker buccal and nasal microarray data to determine whether the genes on the bronchial 314 gene list showed the same direction of change and identified three leading-edge subsets of genes from the bronchial 314 list that were changing expression in the buccal or nasal data in the same direction as seen in the bronchial data. These were a 74 gene subset of genes up-regulated in buccal mucosa of smokers, a 120 gene subset up-regulated in the nasal mucosa of smokers, and a 50 gene subset down-regulated in nasal mucosa. The buccal microarray cel files were downloaded from GEO and analyzed in parallel with the data from the current study (Materials and Methods). Initially, unsupervised hierarchical clustering was performed with the summarized data from the current study, termed SmvsNS, and BuccalCompare for the Sridhar study. Neither dataset showed any pattern of clustering by replicate sample (a vs b) in the case of the SmvNS data, nor by smokers and non-smokers in either dataset.

T-tests comparing the a samples to the b samples in the SmvsNS data were done to evaluate the within-subject variability. There were 871 significant probesets from 53,800, or 1.62%. Comparing smokers to nonsmokers using the same test gave 178 probesets, or 0.33%. A T-test for the BuccalCompare data gave 65 probesets comparing never smokers to smokers and 66 probesets comparing a random grouping of odd numbered arrays against even. Taken together, these results suggest that there is as much or greater variability among subjects than smoking introduces between the two subject types.

SAM (9) and RP (10) were used to develop lists of differentially expressed genes between smokers and nonsmokers. With the SmvNS data, SAM returned 30 significant probesets with a Q value of 0 at a 10% FDR. All 30 probesets were up-regulated in smokers. For the BuccalCompare dataset, there were no significant results from the SAM analysis. With RP analysis, 17 genes were

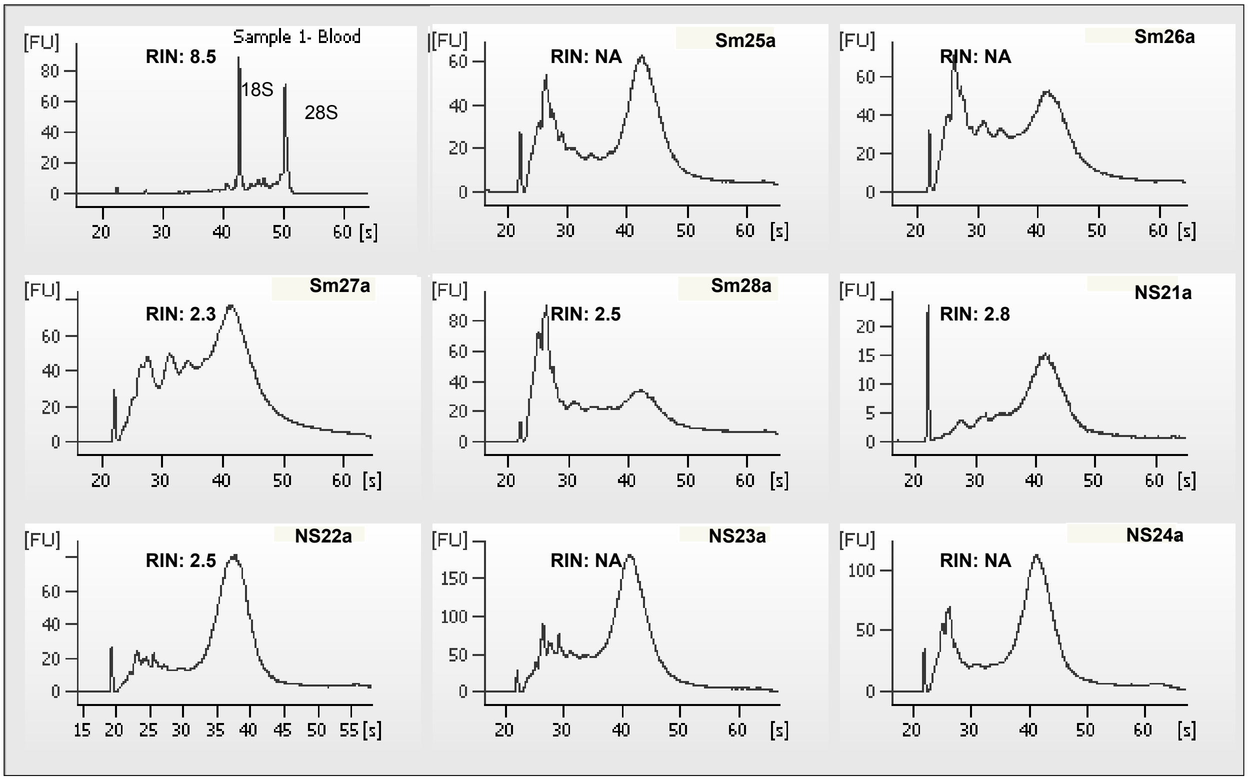


Figure 2A. Buccal mucosa total RNA from smokers and nonsmokers. A. Group a buccal cell samples. Note variation between the isolates in peak heights and species. Sm Smokers, NS nonsmoker. Sample 1, whole blood total RNA, as seen in Figure 1 for comparison, showing 18S and 28S ribosomal peaks. RIN, RNA integrity number, NA RIN not determined.

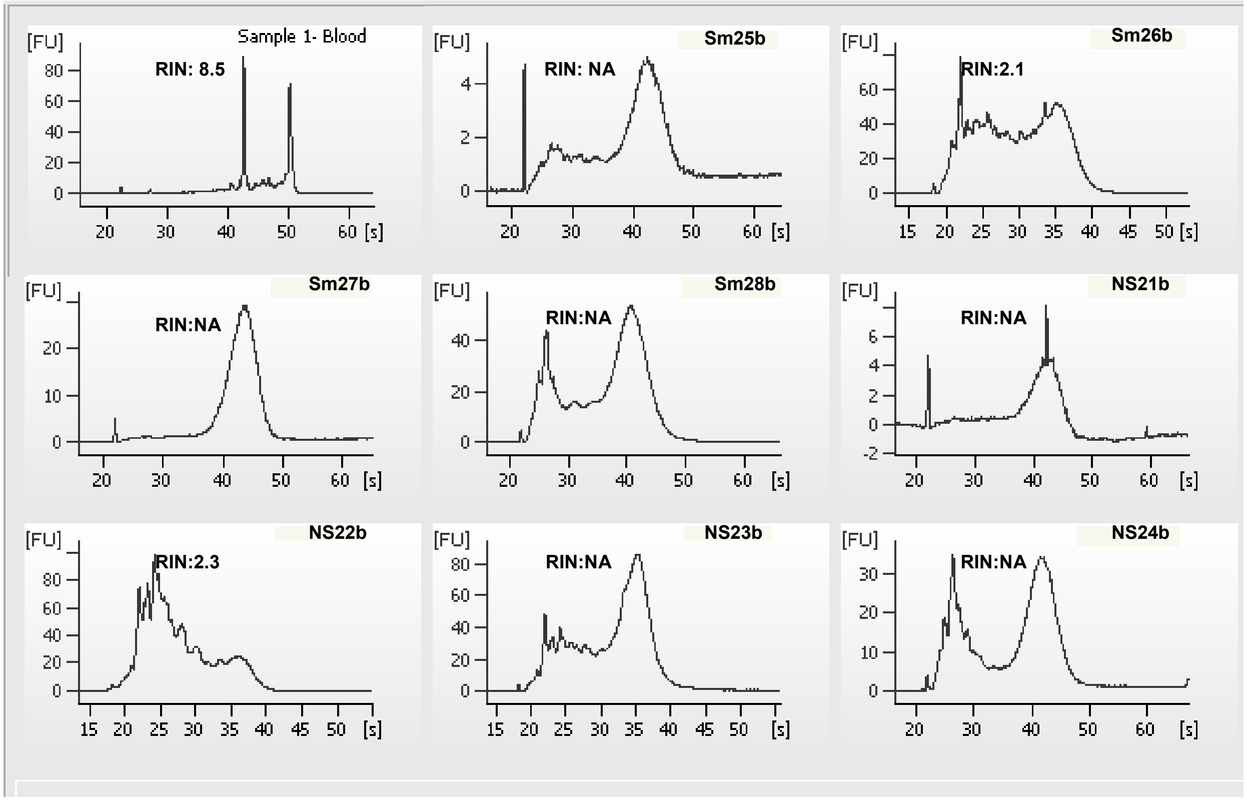


Figure 2B. The group b buccal samples. Sample 1, total RNA from whole blood, is added for comparison. Compare to Fig 2A. For example, Sm26a and Sm26b are from opposite cheeks of same subject and show some similarity in migration pattern. The same variation in peak heights and species between samples is seen here as in Figure 2A. RIN, RNA integrity Number, NA, RIN not determined.

Table 6. Microarray quality metrics.

Samples	Scaling Factor	% Present	NUSE Median	NUSE IQR
Smokers				
25a	24.6444	35.9	0.989	0.021
25b	4.532	47.8	0.991	0.021
26a	19.022	31.8	0.989	0.02
26b	3.451	30	1.013	0.04
27a	**255.647**	**6**	**1.101**	**0.089**
27b	3.713	49.9	0.985	0.021
28a	12.806	22.5	1.027	0.046
28b	4.674	24	1.061	0.073
NonSmokers				
21a	**307.934**	**3.5**	**1.12**	**0.097**
21b	6.852	35	1	0.024
22a	22.185	40.7	0.998	0.019
22b	4.808	47.6	0.993	0.022
23a	20.057	33.6	0.987	0.02
23b	4.689	44.5	0.991	0.022
24a	21.886	39.5	0.988	0.02
24b	3.926	50.9	0.988	0.021

Scaling Factor and % Present were determined with GCOS.
NUSE, [14] and Materials and Methods
IQR, Interquartile range

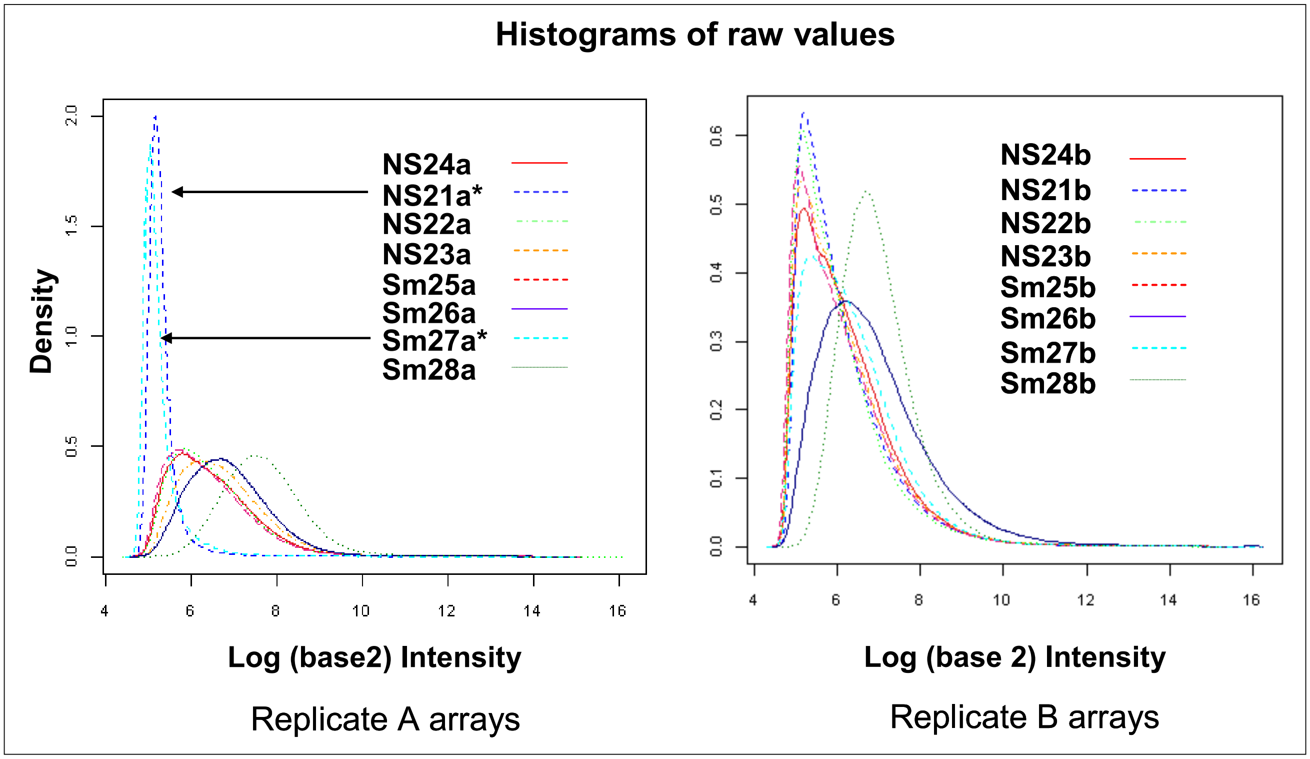


Figure 3. Replicate Samples a and b Raw Signal Value Histograms. Two arrays, NS21a and Sm27a, had low overall signal strength as shown by the intensity plot. The arrays from the matching b cheek NS21b and Sm27b show acceptable values. Note the difference in y-axis density scale. NS nonsmoker, Sm smoker.

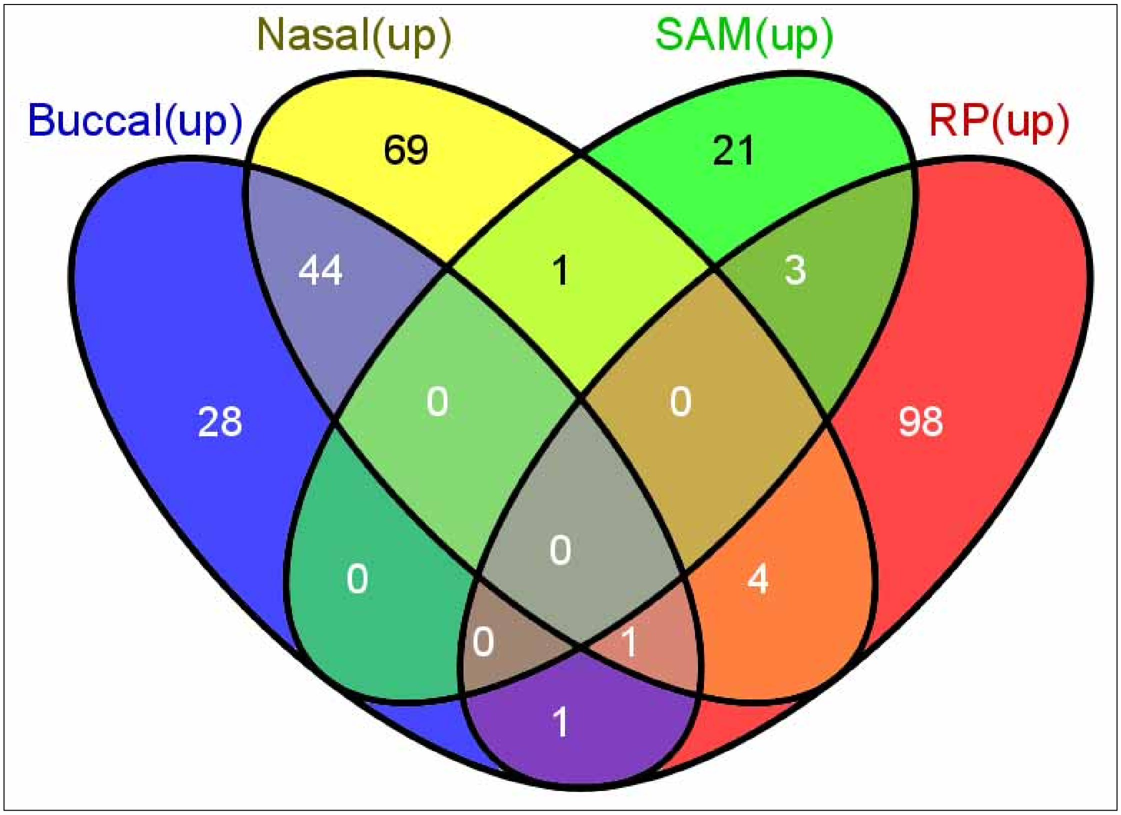


Figure 4. Venn diagram showing overlap between the four gene lists upregulated in smokers.

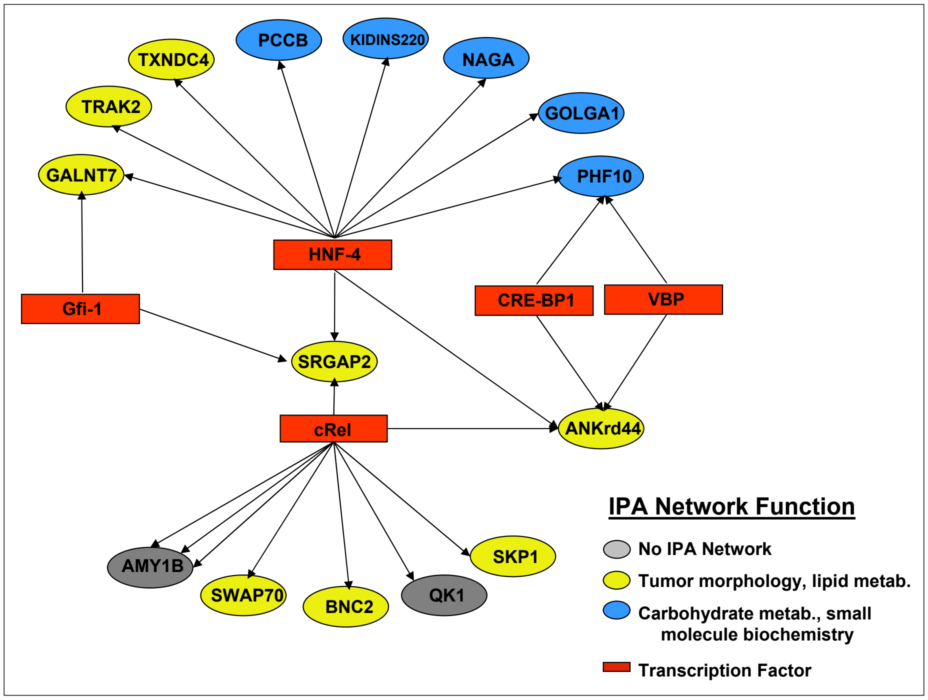
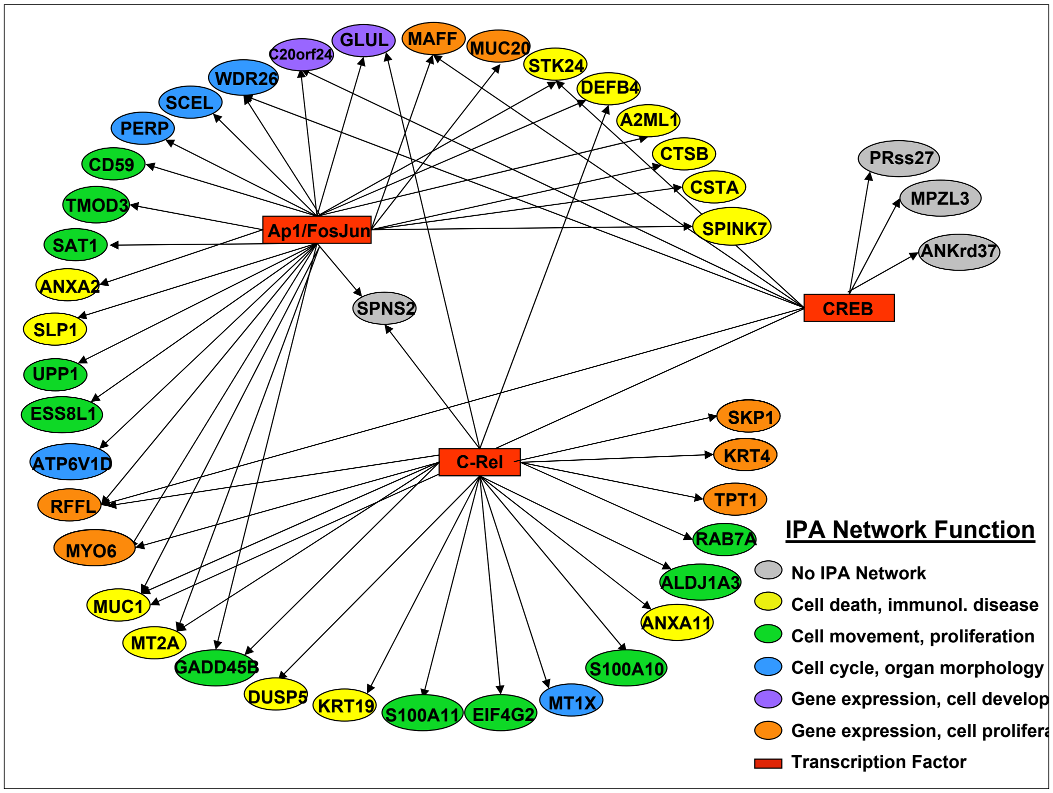


Figure 5. Two graphics showing PAINT TREs with color added to indicate membership in a particular IPA functional network. The ovals represent target genes identified by PAINT as having transcription factor binding sites upstream of the gene. The color of the oval corresponds to the functional networks in which IPA placed the gene. The gray ovals represent genes not included in the IPA network. The rectangles indicate transcription factors. Arrows connect the transcription factors to genes with corresponding upstream binding sites.

A. (Above) Merged results for the SAM_upSm gene list. Twelve of 25 genes are contained in both IPA and PAINT analyses.

B. (Below) Merged results for the RP_upSm gene list. Thirty-eight of 103 target genes are contained in both IPA and PAINT analyses.



found to be down-regulated and 118 genes up-regulated in smokers (Table 7). RP analysis could not be performed on the BuccalCompare dataset since there were no replicates.

Only a few genes were found to be in common between the up-regulated gene lists (Figure 4) (16). The RP_downSm gene list had no overlap with the corresponding Sridhar Nasal_downSm leading edge set. Note that the probesets for the genes on the SAM_upSm and the RP_upSM lists have similar fold change ranges and medians, but probesets in the RP_downSm differed in having overall low signal strength (Tables 8a and b).

Using a similar analysis approach to Sridhar, both the SmvsNS and the BuccalCompare datasets were compared against six gene lists in a GSEA enrichment analysis. The gene lists were the 74 genes in Buccal_upSm , the 120 genes in Nasal_upSm and the 49 genes in Nasal_downSm defined as leading edge subsets by Sridhar ((6), the 25 genes in SAM_upSm, the 107 RP_upSm genes , and the 17 genes in RP_downSm all three lists from the current study (Table 7).

When GSEA analysis of the SmvsNS dataset was performed against all six gene lists, the four lists up-regulated in smokers showed the same expression patterns in the SMvsNS dataset, and the two down-regulated gene lists likewise were down-regulated in the SMvsNS dataset. The same analysis was performed using the BuccalCompare data against the same six gene lists, with the same results. This showed correlation between the SMvsNS and BuccalCompare datasets in terms of the direction of gene expression change for genes in the six sets. However, in the SmvsNS comparison only the SAM_upSm list genes were significantly enriched in the smoker phenotype with FDR q-value 0.029 and p-value 0.025, not the RP_upSm genes. This was unexpected since the RP_upSm gene list was, in fact, derived from the SmvsNS dataset. The BuccalCompare data behaved similarly, with only the Buccal_upSm gene list significantly enriched. This was expected since it was derived from this dataset.

As a check for reproducibility, two subjects (one smoker and one nonsmoker—both cheeks) were retested several months after the initial sampling was performed. Four arrays were generated (11Sm a, b and 12NS a, b). This small dataset was examined with GSEA against the same six gene sets. The results showed that this repeated subset had significant gene enrichment for smokers with the RP_upSm, Nasal_upSm, and Buccal_upSm gene lists with a nominal P-value of 0, an indication of good reproducibility.

Function Analysis

To further evaluate the SmvsNS gene lists for biological coherence, the SAM and RP gene lists were evaluated for over-representation of transcription factor binding sites in the promoters of these genes using the Promoter Analysis and Interactive Tool Set, (PAINT) (12, 17), Materials and Methods, and for shared functional interactions using Ingenuity Pathways Analysis, (IPA Ingenuity IPA version 7.0, Copyright 2009 Ingenuity Systems, Inc., Redwood City CA). Statistically significant transcriptional regulation elements (TREs) were found with 15 of the SAM_upSm and 42 RP_upSm genes. No TREs were found for genes in the RP_downSm genes.

In IPA, 17 of the 25 genes from SAM_upSm could form a single network from two smaller networks sharing broad functional categories including tumor morphology, lipid metabolism, carbohydrate metabolism, and small molecule biochemistry. The RP_downSm genes did not result in any functional networks when examined in IPA. However, 91 of the genes on the RP_upSm list fell into five networks that could be merged into a single large network, indicating shared function. Functional categories for this network included: cell growth, movement, development and death; cell cycle; gene expression, cancer and immunological system development and function.

As a final step in the analysis, genes in TRE networks from PAINT were coded for network function from IPA (Figures 5 a and b). This analysis strongly suggests co-regulation within functional networks and speaks to the transcriptional affects of smoking on buccal cells.

DISCUSSION

This study was focused on determining whether the buccal mucosa could serve as a tissue source for total RNA to be used in relative gene expression studies and biomarker detection by qPCR and microarray analyses. Two previous studies had suggested that buccal cells had efficacy for measuring responses to tobacco smoke exposure (5, 6) and suggested extrapolation of this tissue source to other inhalation or ingestion exposures (5).

Our initial RNA isolations from matched blood and buccal RNA showed a marked difference in the quality of the isolated material between the two sources and showed that there was significant degradation in buccal mucosa RNA. The qPCR results from the matched samples showed an average lower copy number in buccal RNA than blood RNA for all four genes tested and greater variability between subjects (Table 4). The lower copy-number was expected as salivary glands express all four genes at the same or lower level as blood on microarrays; however, the increased variability found between buccal samples over blood is a concern.

The amplification protocols we utilized allowed buccal cell samples to be used in repeated measures experiments, removing the necessity to repeatedly sample to obtain enough RNA for a single microarray. The 50 ng of RNA

Table 7. Gene lists used for the GSEA assay.

Buccal_upSm	Nasal_upSm	Nasal_downSm	SAM_upSm	RP_upSm	RP_downSm
	up in	down in			down in
up in smokers	smokers	smokers	up in smokers	up in smokers	smokers
AACS	AACS	PEX14	AMY1A, 2A	A2ML1	ASMTL
AKR1B1	ABHD2	SIX2	ANKRD44	ACTG1	DMRTC1
AKR1B10	ADH7	TU3A	BNC2	ADM	FLJ33706
ALDH3A1	AKR1B1	PPAP2B	CGGBP1	ALDH1A3	FLJ40243
APLP2	ALDH3A1	ANXA6	FLJ12595	ALG10	H19
ARHE	ANXA3	PECI	GALNT7	ANKRD37	KCNA5
BCL2L13	AP2B1	PDE8B	GOLGA4	ANXA1	LCE1E
BECN1	APLP2	HRIHFB2122	KIDINS220	ANXA11	LOC100131941
C14orf1	ARHE	RUTBC1	LCMT2	ANXA2	LOC158402
CABYR	ARL1	TSAP6	LRMP	ATP6V1D	LOC285708
CAP1	ARPC3	SHARP	NAGA	BCL8	LOC401312
CBLB	AZGP1	BCL11A	PAK2	C14orf129	MEX3D
CCPG1	BECN1	SYNGR1	PCCB	C18orf25	NBPF1,8-11,14-16,20
CDC14B	C14orf1	SEC14L3	PHF10	C20orf24	RGS12
CEACAM5	C1orf8	TLR5	QKI	C4orf7	TMEM107
COPB2	CANX	AK1	RBBP6	CAPN2	TSFM
COX5A	CAP1	AMACR	RNF34	CD59	WNT6
CTSC	CCNG2	LU	SGIP1	CPNE3	
CYP4F11	CCPG1	SERPINI2	SKP1	CRISP3	
CYP4F3	CEACAM5	TNFSF12, 13	SPTBN1	CRNN	
DIAPH2	CEACAM6	TLE2	SRGAP2	CSTA	
DKFZP566E144	CHP	SLIT1	SWAP70	CSTB	
EDEM1	CLDN10	TENS1	TRA2	CTSB	
ENTPD4	COX5A	GGA1	TRAK2	DEFB4	
ERP70	CPNE3	GAS6	TXNDC4	DUSP5	
FLJ13052	CREB3L1	SSH3		ECM1	
FOLH1	CSTA	JAG2		EIF4G2	
GALNT1	CTSC	EPOR		EMP1	
GFPT1	CYP1A1	COL9A2		EPS8L1	
GHITM	CYP1B1	CX3CL1		ERO1L	
GNE	CYP4F11	HNMT		FLG	
GPX2	CYP4F3	C3		FLJ22662	
GSN	DAZ2, 4	FLRT3		FTH1	
GTF3C1	DHRS3	NCOR2		GADD45B	
HIG1	DPYSL3	PCDH7		GLUL	
HTATIP2	DSCR5	SFRS14		GPBP1l1	
JTB	EDEM1	HLF		GPR110	
LAMP2	ERP70	FLJ23514		GRHL1	
LYPLA1	FKBP11	CYFIP2		H3F3A B	
ME1	FKBP1A	FGFR3		HIG2	
MTMR6	FLJ13052	TNS		HOPX	
MUC5AC	FOLH1	FMO2		ITGB1	

The lists in the first three columns are the leading edge gene sets identified in the Sridhar study [6].

The remaining three lists were derived from the current SmvsNS study using SAM or RP analysis as labeled.

(Continued)

Table 7. (Continued) Gene lists used for the GSEA assay.

Buccal_upSm	Nasal_upSm	Nasal_downSm	SAM_upSm	RP_upSm	RP_downSm
	up in	down in			down in
up in smokers	smokers	smokers	up in smokers	up in smokers	smokers
NQO1	FTH1	CUGBP1		KLK10	
PSMB5	GALNT1	ITM2A		KRT13	
PSMD14	GALNT12	SCGB1A1		KRT19	
PTP4A1	GALNT3	TCF7L1		KRT4	
PTS	GALNT7	SLIT2		LCE3D	
RAB11A	GCLM	NFIB		LCN2	
RGC32	GCNT3	KRT15		LGALS3	
RNP24	GFPT1			MAFF	
RPN2	GMDS			MAL	
S100P	GNE			MALAT1	
SEC31L1	GRP58			MALL	
SEPX1	GSN			MPZL3	
SLC35A3	GUK1			MT1F	
SLC3A2	HGD			MT1G	
SLC7A11	HIST1H2BK			MT1H	
SMPDL3A	HMGN4			MT1X	
SORL1	HTATIP2			MT2A	
SPDEF	IDS			MUC1	
SPINT2	IMPA2			MUC20	
SSR4	JTB			MYO6	
TACSTD2	KATNB1			NAMPT	
TALDO1	KIAA0367			NDFIP2	
TARS	KLF4			PADI1	
TIAM1	LAMP2			PER1	
TKT	LOC92482			PERP	
TTC9	LOC92689			PLXNC1	
TXNL1	LRRC5			PPARD	
TXNRD1	LYPLA1			PPL	
UBE2J1	ME1			PPP1CB	
VPS13D	MSMB			PPP1R3C	
WBP5	NKX3-1			PRSS27	
XPOT	NQO1			RAB7A	
	NUDT4			RANBP9	
	P4HB			RFFL	
	PGD			RIOK3	
	PIR			RNF34	
	PLA2G10			S100A10	
	PRDX4			S100A11	
	PTK9			S100P	
	PTP4A1			SAT1	
	RAB11A			SCEL	
	RAB2			SFRS5	

The lists in the first three columns are the leading edge gene sets identified in the Sridhar study [6].
The remaining three lists were derived from the current SmvsNS study using SAM or RP analysis as labeled.
(Continued)

Table 7. (Continued) Gene lists used for the GSEA assay.

Buccal_upSm	Nasal_upSm	Nasal_downSm	SAM_upSm	RP_upSm	RP_downSm
up in smokers	up in smokers	down in smokers	up in smokers	up in smokers	down in smokers
	RAB7			SGIP1	
	RAP1GA1			SKP1	
	RGC32			SLC12A6	
	RNP24			SLPI	
	RPN2			SPINK5	
	S100A10			SPINK7	
	SCGB2A1			SPNS2	
	SCP2			SPRR1B	
	SEC31L1			STK24	
	SEPX1			TACSTD2	
	SLC17A5			TAX1BP3	
	SLC35A1			TMEM49, MIRN21	
	SLC35A3			TMOD3	
	SLC7A11			TMPRSS11B	
	SLC7A11			TMPRSS11E, 11E2	
	SMPDL3A			TncRNA	
	SORL1			TPT1	
	SPDEF			TUFT1	
	TACSTD2			UBAP1	
	TAGLN2			UBB, UBC	
	TCN1			UPP1	
	TIMP1			WDR26	
	TIPARP			ZNF185	
	TKT				
	TLE1				
	TM4SF13				
	TM4SF3				
	TMP21				
	TOM1L1				
	TRA1				
	TTC9				
	TXNDC5				
	UBE2J1				
	UGT1A3, 6				
	UPK1B				
	WBP5				

The lists in the first three columns are the leading edge gene sets identified in the Sridhar study [6].
The remaining three lists were derived from the current SmvsNS study using SAM or RP analysis as labeled.

Table 8a. Fold change of probesets included in the RP and SAM upregulated gene lists.

RP_upSm			
Probeset ID	Signal Strength		
	AveSm	AveNS	Fold change
1560263_at	2839.6	945.0	3.0
1560538_at	307.7	98.4	3.1
1560683_at	1280.0	432.6	3.0
1560684_x_at	1293.2	532.6	2.4
1560712_at	1625.7	349.6	4.7
1564307_a_at	885.7	333.1	2.7
1569603_at	203.8	79.8	2.6
1570233_at	81.4	36.2	2.2
200004_at	338.3	110.4	3.1
200648_s_at	696.7	261.3	2.7
200660_at	717.3	203.3	3.5
200718_s_at	293.8	93.8	3.1
200748_s_at	26221.4	8805.6	3.0
200839_s_at	1941.2	515.8	3.8
200872_at	1805.3	725.3	2.5
200983_x_at	800.6	213.5	3.7
200985_s_at	1693.5	468.0	3.6
201012_at	7017.5	2488.8	2.8
201201_at	16345.2	7844.0	2.1
201324_at	29145.9	16135.9	1.8
201325_s_at	7452.1	3769.0	2.0
201407_s_at	3061.4	1246.2	2.5
201550_x_at	11101.9	4826.8	2.3
201590_x_at	1778.7	428.6	4.2
201650_at	945.3	300.1	3.1
202119_s_at	320.8	94.9	3.4
202129_s_at	1658.8	515.0	3.2
202286_s_at	1550.2	313.3	4.9
202582_s_at	263.2	140.4	1.9
202912_at	1427.0	623.3	2.3
203021_at	2057.7	655.3	3.1
203180_at	484.5	181.8	2.7
203234_at	1313.2	431.8	3.0
203380_x_at	1947.2	663.5	2.9
203407_at	15025.0	4823.6	3.1
203455_s_at	885.6	406.4	2.2
203585_at	568.9	168.3	3.4
204284_at	622.7	172.1	3.6
204326_x_at	1850.8	486.1	3.8
204351_at	1263.1	398.4	3.2
204745_x_at	4730.1	1257.1	3.8
204777_s_at	33731.3	13004.8	2.6
204971_at	20307.1	9778.8	2.1
205064_at	19747.9	6197.7	3.2
205185_at	5303.8	2123.4	2.5
205807_s_at	1933.3	934.6	2.1
206200_s_at	1156.9	360.8	3.2

SAM_upSm			
Probeset ID	Signal Strength		
	AveSm	AveNS	Fold change
1556202_at	235.3	92.4	2.5
1557502_at	564.7	435.0	1.3
1569603_at	203.8	79.8	2.6
1569854_at	406.3	200.6	2.0
200719_at	457.6	259.2	1.8
201567_s_at	804.1	675.2	1.2
202125_s_at	63.5	46.4	1.4
202944_at	109.2	60.1	1.8
204013_s_at	101.8	55.8	1.8
206861_s_at	592.1	254.5	2.3
208498_s_at	382.2	178.5	2.1
208958_at	154.4	66.5	2.3
210369_at	130.3	69.8	1.9
212162_at	319.0	171.1	1.9
216757_at	440.2	259.9	1.7
219126_at	1249.2	867.4	1.4
220716_at	839.0	541.2	1.6
226641_at	602.1	302.8	2.0
227635_at	166.5	95.9	1.7
229942_at	460.7	244.9	1.9
234849_at	103.3	51.9	2.0
235242_at	242.2	140.5	1.7
236288_at	381.5	103.9	3.7
236650_at	1783.5	998.6	1.8
240670_at	575.4	365.5	1.6
242220_at	385.7	186.3	2.1
242299_at	356.7	171.1	2.1
244268_x_at	295.8	126.8	2.3
244348_at	288.2	129.0	2.2
35974_at	39.4	27.7	1.4
	424.5	241.9	1.9
	Average	Average	Median

206461_x_at	4076.3	1131.3	3.6
206471_s_at	3389.5	1431.7	2.4
206884_s_at	8023.2	2087.3	3.8
207356_at	2732.4	785.7	3.5
207802_at	4651.9	1121.0	4.1
207935_s_at	13501.1	6624.8	2.0
208581_x_at	2077.0	524.4	4.0
208683_at	449.6	164.8	2.7
208854_s_at	736.9	218.5	3.4
208855_s_at	2800.6	961.9	2.9
208898_at	753.5	231.4	3.3
208949_s_at	3260.5	1643.8	2.0
209069_s_at	2829.4	992.1	2.9
209154_at	156.9	48.3	3.2
209365_s_at	8156.6	2827.3	2.9
209373_at	1992.7	631.5	3.2
209457_at	1721.0	759.6	2.3
209792_s_at	255.8	69.1	3.7
210427_x_at	1533.4	368.6	4.2
210480_s_at	909.4	322.8	2.8
210592_s_at	1631.9	600.0	2.7
211296_x_at	2773.5	1111.9	2.5
211597_s_at	2493.3	888.5	2.8
211945_s_at	381.4	99.7	3.8
211960_s_at	89.0	34.4	2.6
211970_x_at	12242.4	5876.6	2.1
211983_x_at	30559.4	13369.9	2.3
212185_x_at	6117.9	2063.0	3.0
212266_s_at	2409.3	647.6	3.7
212284_x_at	13069.8	5894.7	2.2
212531_at	1313.8	283.7	4.6
213240_s_at	10862.7	4773.7	2.3
213503_x_at	1746.2	452.3	3.9
213560_at	1431.4	468.8	3.1
213693_s_at	6514.1	2001.7	3.3
214399_s_at	24445.3	12004.8	2.0
214657_s_at	401.5	163.4	2.5
215704_at	753.0	272.4	2.8
217165_x_at	1979.5	454.2	4.4
217508_s_at	251.1	90.1	2.8
217739_s_at	761.8	292.3	2.6
217835_x_at	651.7	277.9	2.3
218107_at	339.6	112.3	3.0
218454_at	696.3	220.9	3.2
218507_at	586.8	198.5	3.0
218779_x_at	1644.2	365.1	4.5
220090_at	13668.3	4995.6	2.7
220431_at	11097.6	4104.3	2.7
220990_s_at	308.5	97.7	3.2
221655_x_at	1918.8	394.5	4.9
221665_s_at	1156.7	265.7	4.4
222392_x_at	2353.1	1000.3	2.4

222646_s_at	8991.7	3433.9	2.6
222830_at	199.0	57.5	3.5
223077_at	330.1	93.6	3.5
223239_at	118.7	33.2	3.6
223596_at	64.7	21.8	3.0
223720_at	1360.4	602.1	2.3
223739_at	1066.5	279.9	3.8
224328_s_at	989.6	359.8	2.8
224565_at	2330.1	980.9	2.4
224566_at	1424.5	471.3	3.0
224567_x_at	12920.9	7753.9	1.7
224585_x_at	15711.4	6478.1	2.4
224799_at	821.4	269.1	3.1
225671_at	570.2	189.0	3.0
225750_at	1694.4	622.7	2.7
226622_at	840.0	251.9	3.3
226675_s_at	11677.6	8529.3	1.4
227337_at	1919.1	823.5	2.3
227747_at	4257.5	1262.4	3.4
229152_at	690.4	166.6	4.1
231735_s_at	31993.2	13536.9	2.4
232056_at	220.4	85.8	2.6
232074_at	675.6	226.4	3.0
233513_at	215.3	79.3	2.7
234989_at	15563.5	4119.7	3.8
236288_at	381.5	103.9	3.7
237919_at	360.7	164.0	2.2
238689_at	258.8	92.8	2.8
244677_at	940.0	271.3	3.5
36711_at	1626.5	463.2	3.5
37152_at	999.2	304.9	3.3
46270_at	527.3	154.8	3.4
91826_at	3230.9	1056.8	3.1
	4421.6	1781.3	3.0
	Average	Average	Median

Table 8b. Fold change of probesets included in the RP down-regulated gene list.

RP_downSm Probeset ID	Signal Strength			
	AveSm	AveNS	Fold change	1/fold change
1553998_at	19.5	57.5	0.3	2.9
1555032_at	190.7	413.2	0.5	2.2
1556721_at	24.3	64.4	0.4	2.7
1557818_x_at	149.8	334.7	0.4	2.2
1559224_at	185.0	431.1	0.4	2.3
1560520_at	30.4	71.7	0.4	2.4
1561061_at	87.7	199.2	0.4	2.3
1564281_at	23.0	66.6	0.3	2.9
1564996_at	86.4	280.3	0.3	3.2
1566999_at	45.1	99.3	0.5	2.2
1567139_at	53.7	127.4	0.4	2.4
1567697_at	17.5	42.5	0.4	2.4
1568365_at	73.1	188.6	0.4	2.6
206762_at	72.5	209.9	0.3	2.9
214331_at	37.1	95.6	0.4	2.6
224495_at	21.4	53.0	0.4	2.5
224997_x_at	96.4	419.1	0.2	4.3
227926_s_at	53.2	121.4	0.4	2.3
233891_at	207.2	450.6	0.5	2.2
236769_at	69.1	151.6	0.5	2.2
240411_at	139.9	322.6	0.4	2.3
36554_at	101.2	230.4	0.4	2.3
71933_at	88.4	318.6	0.3	3.6
91816_f_at	39.8	118.5	0.3	3.0
	79.7	202.8	0.4	2.4
	Average	Average		Median

we used for amplification can routinely be isolated from a single swab (Table 1), in contrast to the 8 ug required by Sridhar et al. (6), which was pooled from multiple sampling of the same individual over a period of 6 weeks. In most cases, the array quality was acceptable, but with buccal RNA, arrays did have a higher failure rate than is typical for arrays hybridized with target material from blood RNA. Two samples from 16 failed in hybridization, where matching samples from the other cheek passed. This opens the possibility that samples from both cheeks would be required to insure that every sample was collected in a study, but we found the intra-subject variability to be high as well. The availability of the Sridhar buccal dataset provided comparison data and, along with the previous work from this group (7), provided published lists of genes from buccal and nasal cells that change expression levels due to smoking. Gene lists developed from the current study did not overlap extensively with each other or with the Sridhar lists. However, by using the independent analysis tools, PAINT and IPA, a cohesive function/cotranscription network was generated, suggesting two non-random sets of genes upregulated in smokers. TFBS analysis is a good complement to a functional analysis such as IPA because it has no *a priori* assumptions about gene function, relying instead on promoter sequence alone. The analysis results suggested that using an approach that included these two complementary methods is useful for evaluating candidate genes.

The analysis conducted with GSEA was significant because there was perfect concordance between gene lists derived from each of the two datasets for the direction of change in expression between smokers and non-smokers. The results from the small repeated dataset were an indication of reproducibility with this system. This validated the methods used in the current study to discover differentially expressed genes. However, the lack of consistent, statistically significant enrichment for the smoker phenotype with GSEA analysis, the within-subject variability of RNA quality, and degradation in RNA derived from buccal cells highlight the difficulties to be expected when using buccal-cell RNA for differential expression testing.

CONCLUSIONS

This study was a straightforward evaluation of buccal mucosa as a tissue useful for evaluating relative gene expression changes using an analysis scheme containing well-validated and commonly used analysis tools. Isolation and amplification techniques were successfully modified from those used with whole blood. The level of degradation found was not unexpected, and we were able to successfully perform qPCR with the buccal RNA. Somewhat surprising was that, given the poor quality of the RNA, the quality of the majority of the microarrays was acceptable and that several lists of genes changing expression in smokers, compared to nonsmokers, resulted from statistical analysis of the arrays. There was evidence of reproducibility in expression change, but the borderline significance level of the lists questions the validity of the findings. Therefore, using buccal tissue RNA with 3' amplification may be a suitable tissue choice and preparation approach when assaying specific, highly differentially expressed gene targets that could overcome the limitations of subject variability and sample degradation. However, our findings suggest that this may be a difficult tissue to use, requiring replicate sampling and arrays, and possibly a different technology such as an array format or amplification method designed for heavily degraded template material.

METHODS

Sample Collection

All sample collection was performed with the informed consent of the study participants under the auspices of the local IRB. Blood samples were collected in PAXgene™ Blood RNA tubes (PreAnalytix/Qiagen; Valencia, CA) according to the manufacturer's published protocol. Urine samples for nicotine and cotinine testing were collected in urine cups without preservative and refrigerated until shipping to a clinical lab (Diagnostic Laboratory of Oklahoma; Oklahoma City, OK). All nonsmokers were below the level of detection for both nicotine (10 ng/ml) and cotinine (40 ng/ml). All smokers were greater than 500 ng/ml for nicotine and 900 ng/ml for cotinine. The expected levels for smokers are a concentration greater than 100 ng/ml for nicotine and 200 ng/ml for cotinine.

Buccal samples were collected using sterile Cytobrush Plus® (Medscand Medical; Guttenberg, NJ). Subjects were asked not to eat for the 30 minutes prior to sampling and rinsed their mouths with a minimum of 20mL of water before sample collection. Two buccal samples were collected from each subject and processed separately as either "a" or "b" samples. Cheeks were brushed for 30 seconds, then brushes were plunged into 2-mL tubes containing 1.0 ml of RNA-Later (Invitrogen; Carlsbad, CA). The brush ends were cut off with sterile surgical scissors such that the 2-ml tubes could be capped. RNA was purified from buccal cell swabs immediately after collection.

RNA Purification

RNA isolation from blood samples was performed according to the protocol in the PAXgene™ Blood RNA Purification Kit (18) with the optional on-column DNase treatment. A blood total RNA control sample was created by pooling purified RNA samples from three individuals not participating in either study.

Buccal-cell RNA was purified using the RNeasy Micro Kit (Qiagen; Valencia CA) with the modifications found in Spivack et al. (5) and here. Cells were pelleted by centrifugation at 4,000 X g. The brush was removed from the tube by scraping the bristles against the lip of the tube to remove any adhered cells and the pellet reformed by centrifugation as above. RNAlater was pipetted off the pellet and the pellet washed with ice-cold PBS and the PBS removed after centrifugation, as above. Two microliters of polyC (Sigma Chemical; St. Louis, MO) and 350 ul Buffer RLT (RNeasy Micro Kit) containing 10ul/ml beta-mercaptoethanol was added and the pellet passed through a 25 ga needle to lyse the cells. The lysate was centrifuged at 20,000 xg for 3 minutes and the supernatant transferred to a fresh microfuge tube. Then 350 ul 70% ethanol was added, mixed well by pipetting, and the sample applied to a MinElute column (RNeasy Micro Kit) and centrifuged at 8000 xg for 30 seconds. The column was washed twice with 350 ul of RW1 buffer (RNeasy Kit) followed by centrifugation at 8000 xg for 15 seconds. The column was placed in a fresh 2 ml collection tube and 500ul RPE buffer (RNeasy Micro Kit) was added. The column was centrifuged at 8000xg for 30 seconds; 500 ul of freshly prepared 80% ethanol was added to the column followed by centrifugation for 2 minutes at 8000 xg. The column was transferred to a fresh 2 ml collection tube, with the cap open, and centrifuged at 16,000 xg for 5 minutes. The RNA was eluted by adding 30ul prewarmed (50-55° C) RNase-free water to the membrane. After 2 minutes incubation, the column was centrifuged at 16,000 xg for 2 minutes. Spectrophotometric analysis showed a large 230nM component, potentially salt carryover. To reduce this, the RNeasy Micro Kit protocol for RNA cleanup and concentration (December 2007 version) was used as written by the manufacturer for sample volumes less than 100 ul.

RNA quality was assessed from Agilent Bioanalyzer 2100 (Agilent; Santa Clara, CA) traces using the Agilent RNA 6000 Nano Series II kit following manufacturer's directions with 1 ul of sample. Yield was determined on a Nanodrop 1000 spectrophotometer (Thermo Scientific; Waltham, MA) (Table 1).

qPCR. Primers for qPCR were designed using Beacon Designer 7.0 (PREMIER Biosoft International; Palo Alto, CA). Primers were synthesized and HPLC purified (Integrated DNA Technologies; Coralville IA). For three genes, ITAG5, ANKRD28, and TMEM8, three sets of primers were designed to span the mRNA. See Table 3 for the sequences, positions, of the primer sets on the respective transcript, concentrations, and annealing temperatures.

Template material for qPCR was prepared from 50 ng aliquots of total RNA that were reversed transcribed and amplified using either the WT-Ovation™ Pico System or the Ovation™ RNA Amplification System V2, #3300, 3100, respectively (Nugen Technologies, Inc.; San Carlos, CA). All qPCR reactions were 25 ul and performed in triplicate with a SYBR® green based assay, PerfeCta SYBR Green FastMix, Low ROX, #95074-05k (Quanta Biosciences; Gaithersburg, MD) using 1 ng of amplified template material per reaction except in the amplification comparison series, where 5 ng/reaction was used. Cycling was performed on a Stratagene MX3005p (Agilent Technologies; La Jolla, CA). Cycling parameters were one cycle of 2 min 95° C, 40 cycles of 15 sec 95° C, 30 sec optimum annealing temperature, 15 sec 72° C extension, then a dissociation curve with 1 min 95° C, 30 sec at optimum annealing temperature, and dissociation ramp rate at 0.01 degree/sec with all points data collection on. qPCR data were analyzed using qBase version 1.3.5 (19). qPCR product size was assessed with Agilent DNA 1000 Series II (Agilent Technologies) microfluidics chips.

Microarray target preparation. For microarray target material, 50 ng total RNA was reverse transcribed and amplified, per the manufacturer's protocols using the Ovation™ RNA Amplification System V2 (Nugen Technologies, Inc.), fragmented and biotin-labeled, using the FL-Ovation™ cDNA Biotin Module V2, #4200 (Nugen Technologies, Inc.). Gene expression was determined by hybridization of the labelled template to hgU133 Plus 2.0 human microarrays (Affymetrix, Inc.; Santa Clara, CA). Hybridization cocktail synthesis and post-hybridization processing was performed according to the "Affymetrix GeneChip Eukaryotic Array Analysis" protocol found in the appendix of the protocol book for the fragmentation kit. Arrays were hybridized for 18 hours and washed using fluidics protocol FS450_0004 on a GeneChip Fluidic Station 450 (Affymetrix, Inc.).

Microarray pre-processing. Quality assessment of the arrays was performed with the tools available in the Gene Chip Operating Software, version 1.4 (Affymetrix, Inc.) and the Bioconductor packages AffyQCReport (20) and AffyPLM (21), R version 2.8, Bioconductor version 2.3 (22). The microarray data have been assigned series number GSE16149 in the Gene Expression Omnibus (GEO).

Microarray data analysis. Array data were processed with Robust Multiarray Average (RMA) (23) using the package available at the Automated Microarray Pipeline (AMP) (24) and quantile normalized. Differential expression analysis comparing smokers to nonsmokers was performed with both Significance Analysis of Microarrays, SAM, (9) and Rank Product Analysis, RP, (10). For RP analysis, the samples matching the two poor quality arrays were removed, as this analysis utilizes the ranked expression values from replicate samples. This left 12 arrays, six in each replicate group, a and b, for this analysis. Unsupervised hierarchical clustering, T-tests, SAM and RP were performed using the packages available on the MultiExperiment Viewer, version 4.3.01 (MeV) (25, 26) with default settings. Gene Set Enrichment Analysis, GSEA version 2.04 (8, 27), was used to test the array data for enrichment of differentially expressed genes. The default settings were used, except the minimum size for gene sets was decreased to ten to allow analysis against the RP_downSm list, which GSEA reduced from 17. The same microarray differential expression analysis pipeline was used on the data from series GSE8987 from the GEO database (6), which were designated mouth and current smokers.

The output gene lists of differentially expressed genes from RP and SAM were evaluated for biological significance using Ingenuity Pathway Analysis, IPA, (Ingenuity Systems, Inc.; Redwood City, CA) for a core analysis. PAINT, promoter analysis and interaction network version 3.6 (12) analysis using the TRANSFAC public database (28), was used with the same gene lists examining both strands to 2000 bases upstream looking for transcription factor binding sites and summing in TREs any potentially co-regulated genes.

REFERENCES

1. Thompson MD, Bowen RA, Wong BY, et al.: Whole genome amplification of buccal cell DNA: genotyping concordance before and after multiple displacement amplification. *Clinical Chemical Laboratory Medicine* 2005, 43(2):157-62.

2. Ceder O, van Dijken J, Ericson T, Kollberg H: Ribonuclease in different types of saliva from cystic fibrosis patients. *Acta Paediatrica Scandinavica* 1985, 74:102-6.

3. Vondracek M, Xi Z, Larsson P, et al.: Cytochrome P450 expression and related metabolism in human buccal mucosa. *Carcinogenesis* 2001, 22(3):481-8.

4. Spira A, Beane J, Schembri F, et al.: Noninvasive method for obtaining RNA from buccal mucosa epithelial cells for gene expression profiling. *BioTechniques* 2004, 36:484-7.

5. Spivack SD, Hurteau GJ, Jain R, et al.: Gene-environment interaction signatures by quantitative mRNA profiling in exfoliated buccal mucosal cells. *Cancer Research* 2004, 64:6805-6813.

6. Sridhar S, Schembri F, Zeskind J, et al.: Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. *BMC Genomics* 2008, 9:259-.

7. Shah V, Sridhar S, Beane J, et al.: SIEGE: Smoking induced epithelial gene expression database. *Nucleic Acids Research* 2005, 33(database):D573-9.

8. Subramanian A, Tamayo P, Mootha VK, et al.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 2005, 102(43):15545-50.

9. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 2001, 98(9):5116-21.

10. Breitling R, Armengaud P, Amtmann A, Herzyk P: Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters* 2004, 573:83-92.

11. Ingenuity Pathway Analysis (IPA) [http://www.ingenuity.com/] (access date 09.25.09).

12. Vadigepalli R, Chakravarthula P, Zak DE, et al.: PAINT: A promoter analysis and interaction network generation tool for gene regulatory network identification. *OMICS* 2003, 7(3):235-52.

13. SymAtlas/BioGPS gene atlas database [http://biogps.gnf.org] (access date 09.25.09).

14. Bolstad BM, Collin F, Brettschneider J, et al.: Quality assessment of Affymetrix GeneChip data. Bioinformatics and computational biology solutions using R and Bioconductor. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Gentleman R, Carey V, Huber W, Irizarry RA, Dudoit S eds.: New York, Springer; 2005: 33-47.

15. SIEGE-Smoking induced epithelial gene expression database [http://pulm.bumc.bu.edu/siegeDB/].

16. Oliveros JC: VENNY. An interactive tool for comparing lists with Venn Diagrams [http://bioinfogp.cnb.csic.es/tools/venny/index.html] (access date 09.25.09).

17. PAINT: Promoter analysis and interaction network toolset (V 3.6) [http://www.dbi.tju.edu/dbi/tools/paint/] (access date 09.25.09).

18. PreAnalytiX Technical Notes [http://www.preanalytix.com/Tech_Notes.asp#] (access date 09.25.09).

19. Hellemans J, Mortier G, De Paepe A, et al.: qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biology* 2007, 8(2):R19.

20. Parman C, Hallman C, Gentleman R: AffyQCreport [http://www.bioconductor.org/packages/2.4/bioc/html/affyQCReport.html] (access date 09.25.09).

21. Bolstad BM: affyPLM [http://www.bioconductor.org/packages/bioc/1.6/src/contrib/html/affyPLM.html] (access date 09.25.09).

22. Bioconductor: open source software for bioinformatics [http://www.bioconductor.org/] (access date 09.25.09).

23. Irizarry RA, Hobbs B, Collin F, et al.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003, 4(2):249-64.

24. AMP: Automated Microarray Pipeline [http://compbio.dfci.harvard.edu/amp] (access date 09.25.09).

25. TM4 Microarray Software Suite [http://www.tm4.org] (access date 09.25.09).

26. Saeed AI, Sharov V, White J, et al.: TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 2003, 34(2):374-8.

27. Gene Set Enrichment Analysis [http://www.broad.mit.edu/gsea] (access date 09.25.09).

28. TRANSFAC database of conserved transcription factor binding sites [http://www.gene-regulation.com/pub/databases.html] (access date 09.25.09).